WebTheme: Visual Text Mining for the World-Wide Web

Mark Whiting, Lucy Nowell, Kevin Walker
Pacific Northwest National Laboratory[1]
PO Box 999
Richland, WA 99352

Abstract - WebTheme™ is an interactive, graphical application that supports visualization and analysis of World Wide Web documents. WebTheme enables users to explore themes and concepts found among Web documents by graphically depicting the contents of large collections of Web pages. This system supports automated information harvesting, user-specified filtering, and visualization of Web page contents, using advanced statistical techniques. WebTheme can be used to characterize and analyze one or more Web sites or to visually organize and examine Web search results. Users interacting with these visualizations can rapidly identify themes and concepts found within the text of the pages and choose specific areas of the Web information they wish to explore.

I. INTRODUCTION

Scientists and engineers are discovering the Web as a convenient repository for casually and formally published information. The Web is now often the primary reference for many information collection activities. Web information grows rapidly, changes frequently, and can be well-advertised and apparent to users or inconspicuous and hidden within the depths of a Web site. For an information user, it is often difficult to develop an overall understanding of a site or discover the most interesting nuggets of information without extensive and time-consuming manual processing.

WebTheme provides an alternative that helps users see patterns in large collections of Web pages and make choices about which documents to read. Starting from either a list of URLs or a query string, WebTheme harvests Web pages and then automatically organizes and graphically depicts their contents. WebTheme organizes the results visually, showing how retrieved Web pages relate to one another. Users interacting with these visualizations can rapidly identify themes and concepts found within the text of the pages and choose which areas of the Web they wish to explore further.

II. RELATED WORK

Several systems exploit the advantages of information visualization and information retrieval to certain extents. Many visualization interfaces for information retrieval systems present ranked query-document similarity and clustering. VIBE [1] allows users to input query terms, which are associated with a portion of the visualization window, with document icons positioned to illustrate the relevance of documents to the selected terms. TileBars [2] developed at Xerox PARC allows the user to enter search terms as topics. After the system retrieves documents, a graphical, tiled bar is displayed next to the title of each document showing the relationship between the document and query terms. Other efforts have focused on creating maps of Web site content. Mappucino allows visual mapping and exploration of web sites [3].

III. BACKGROUND

WebTheme was created as an internal research and development project by the Pacific Northwest National Laboratory (PNNL) in 1996. The goal was to demonstrate the feasibility of applying advanced visualization techniques to Web data. Subsequent sponsorship from the NASA Goddard Space Flight Center advanced WebTheme from proof-of-concept to the current prototype version. WebTheme has been designed to enhance the effectiveness of user access to NASA textual data by supporting hands-on analysis through advanced text visualization techniques. Targeted users of WebTheme include professional analysts, scientists, engineers, and educators.

IV. WEBTHEME DESIGN

WebTheme was designed to leverage sophisticated text analysis algorithms implemented on UNIX server systems and to provide flexible, interactive visualizations to client desktop systems via a Web browser. WebTheme is based on PNNL's Spatial Paradigm for Information Retrieval and Exploration (SPIRE) system [4], which produces advanced visualizations from user-provided data sets. SPIRE runs on a UNIX computer workstation with pre-captured data sets. Like SPIRE, WebTheme uses PNNL's text analysis engine, running on a UNIX server. The text engine applies advanced statistical methods to identify the key topics within a document set and produces a document vector, or numerical representation, of each document's essence as it relates to other documents in the set. Then, the document vectors are used by a projection algorithm to produce a two-dimensional numerical representation that can be plotted on a computer screen to create a Galaxies display and a ThemeView™ visualization. WebTheme delivers additional

[1] Pacific Northwest National Laboratory is managed for the U.S. Department of Energy by Battelle Memorial Institute under Contract DE-AC06-76RL0-1830.

capabilities through text analysis and visualization for Web data to client software on desktop platforms.

The WebTheme system architecture is shown in Figure 1. WebTheme harvests documents from the Internet, processes them on the WebTheme server that hosts the SPIRE text analysis processes, and delivers visualizations and analysis tool capabilities to users via a Web server. So, although SPIRE and the WebTheme server run on a UNIX computer,

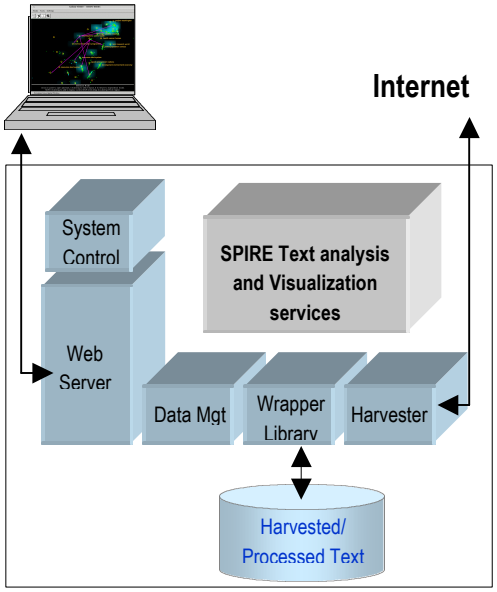

Figure 1: WebTheme Architecture

WebTheme can be accessed by Windows and Macintosh computers running Netscape or Internet Explorer.

The WebTheme user interface provides an environment where the user can initiate a harvest of Web documents and specify parameters for the harvest and processing of retrieved documents. Also, the user interface supports visualization of the results through a set of interactive tools that provide a visual representation of the document space, showing thematic relationships between documents pictorially. Java applets interface with the prepared data on the server, allowing the user to query and probe the visual space and groups of documents. Individual documents or sets of documents can be retrieved from the server and examined. In addition, the original Web page that produced a document can be retrieved and displayed in a separate browser window.

Users may specify Web-based retrieval in any one of three ways. First, an anchor URL may be specified, from which WebTheme agents will "spider" down into the site to retrieve pages. Second, the user may specify a search engine query, to be sent to either Alta Vista, Google, or Excite. Documents resulting from that search will then be retrieved and processed by the text engine. Third, a user may specify a Z39.50 protocol retrieval from a digital library. This functions more like a database query than the other two approaches. Users may set several other parameters that allow them to:

- Block harvesting of certain Web sites that are known to be of no interest.
- Limit the search to a particular Internet domain.
- Specify how many layers or levels of links should be followed from each page on the initial list.
- Specify minimum and maximum numbers of pages to harvest.
- Specify that the search should proceed for a specified period of time, rather than setting a target number of documents to be retrieved.
- Set filters to eliminate certain kinds of unwanted items, including those in foreign languages.

In the case of the URL-based or search engine query, the harvester behaves as a specialized Web client, making contact with Web servers and then requesting and receiving Web documents. Harvesting requests are generated in parallel from the WebTheme server. This significantly increases harvest speed, because each request may involve delays from the remote servers. The harvesting process retrieves the documents in the initial list, searches those documents for HTML links, and continues by following links on retrieved documents to whatever depth the user requested. Harvesting continues until a user-specified number of documents are retrieved or a user-specified time period has elapsed.

For the example in this paper, we specified a Google query consisting of the terms, "genetically", "engineered", and "irradiated". An initial review of the results of such a Google query looked promising, but the 63 pages of results were too numerous to work with without additional help.

Figure 2 shows a WebTheme visualization called a ThemeView.

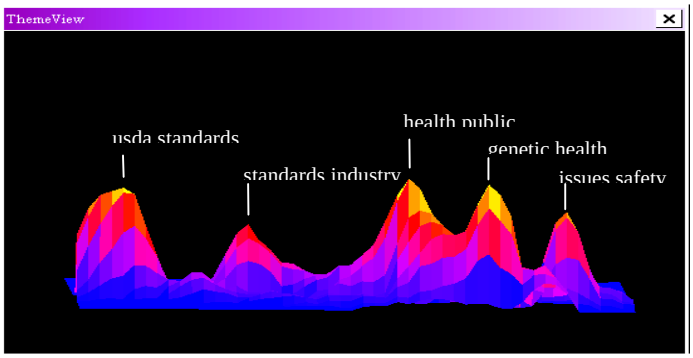

Figure 2: ThemeView

This visualization shows the results of Google query harvest. We see strong themes related to health, safety, USDA, and standards. Note that words appearing together as a peak label are not processed as a phrase; they are simply terms that are both strongly evident at that point in

the collection, not necessarily in the same documents.

The WebTheme Galaxies visualization for the same data set is shown in Figure 3. Each green point in the visualization

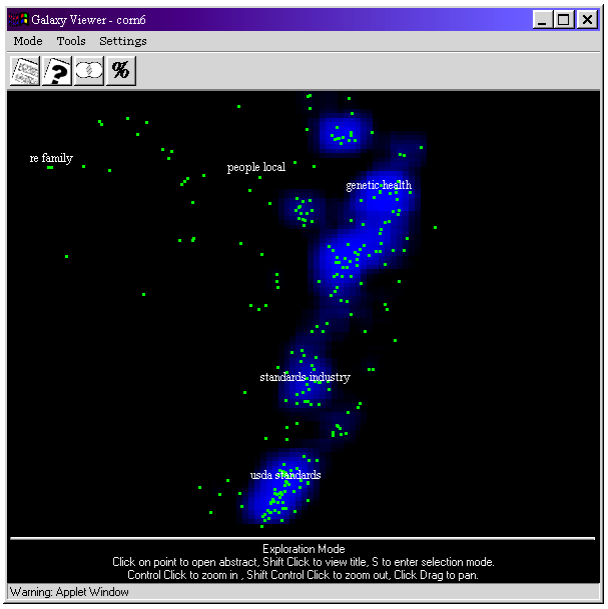

Figure 3: Galaxies

represents the text component of an individual Web page that was harvested. The distance between points indicates their thematic similarity. Thus, if points in the visualization are close together, then it is likely that the corresponding documents will contain thematically similar information. If they are far apart, the documents will probably be very different. The spatial layout of the points in the X-Y plane is not meaningful to users.

Areas of the visualization that contain high thematic content are highlighted in blue. These blue ThemeClouds, resembling nebulae, are labeled to indicate where significant topics occur within the document set. Labels are automatically shown for peaks, or locations of greatest density, to provide some orientation to the set.

The Galaxies and ThemeClouds visualizations are rich sources of insight into the thematic content of the collection. By examining ThemeCloud labels, users can rapidly become familiar with the general topics and themes represented. Sometimes, however, users want to know more about the thematic content of a region than the ThemeClouds labels provide.

Users may further explore the set using a variety of tools. At the simplest level, document titles may be revealed individually or in groups. The Probe Tool allows users to explore the composition of a ThemeCloud from the inside out, showing the list of themes associated with any location on the screen, whether or not a dot is in that location. The list is ordered from the strongest to weakest thematic significance, indicated by the numbers to the right of the Probe Data window.

The Gisting Tool lists frequently occurring terms in a group of selected documents, reporting the number of documents in that set which contain each word. The list is ordered from highest frequency to lowest. Gisting a group of 52 selected documents from the region at the bottom of the Galaxies visualization shown in Figure 3 reveals that the word "food" and "agriculture" occurs in 48 of those documents, "farmer" occurs in 40. This is not surprising, given the theme label "usda standards".

The WebTheme Link Mode allows users to see which documents in the harvested set are hyperlinked to one another. This feature is especially useful for a set that has been harvested from a list of URLs, with multiple levels of links followed. Showing the links allows the user to see the link structure as well as the thematic structure of the Web site. It may also revealed patterns of citation in a collection harvested from a query. In Figure 4 these links are shown as bright rose lines between the dots for linked items. The

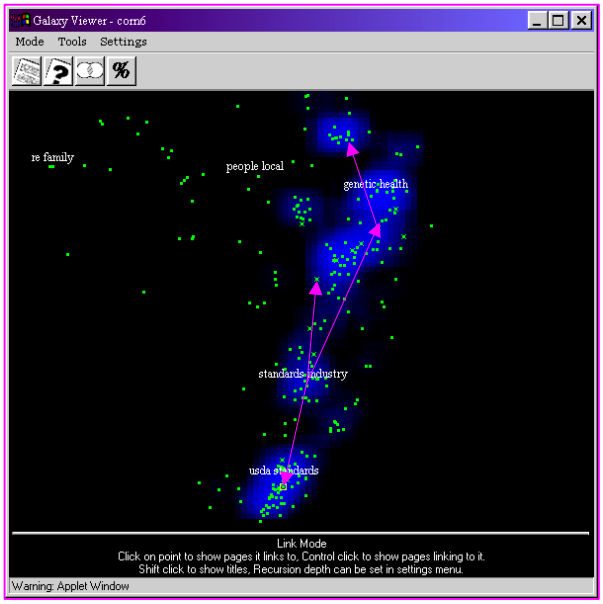

Figure 4: Galaxies visualization with Links

source of the links in the figure is a news article on organic foods, with links to other articles on food labeling and standards. WebTheme includes the capability to search the harvested collection using two types of Query Tool searches: Words in Document and Query-by-Example. The Query Tool window is shown in Figure 5, along with the Group Tool; the Galaxies visualization has been zoomed in on the upper ThemeCloud.

The Words query operation is Boolean, selecting a set of dots for documents that contain the words in the query. If the Document Viewer is opened, titles matching the selected documents appear in the top of the viewer. Users

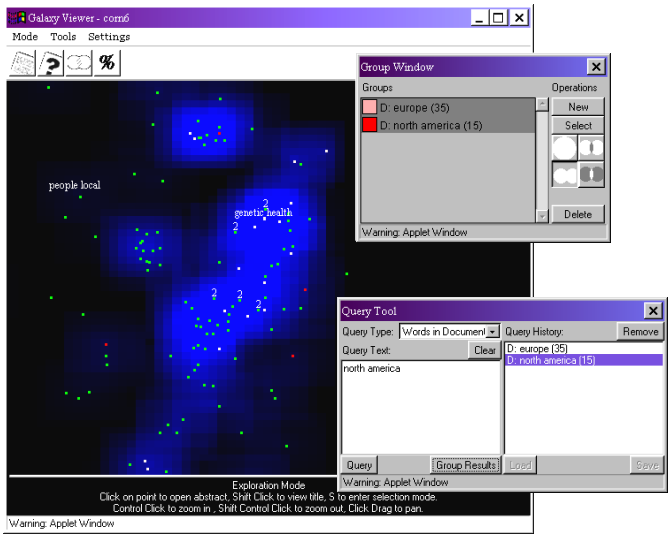

Figure 5: Query and Group Tools

also have the option to open documents in their Web browser.

Query-by-Example triggers a vector space search and selects the document that is the best match to the query — the one that is closest in the n-dimensional vector space to the query vector. A slider on the Query Tool window allows the user to vary the number of documents selected. By manipulating the slider and thus changing which dots are highlighted, the user can distinguish the location of documents that are closest to the query in the vector space from those that are further away.

A Query History pane is provided on the right side of the Query Tool window. The Query History is a line-by-line record of queries made during a search session. Each line of the query history represents a single query. The first letter on the line signifies the type of query, followed by a colon and the first few words of the query string. The number of documents retrieved is shown in parentheses at the end of the line. The most recently executed query will be highlighted in the Query History. In Figure 5, we see that two Words in Document search results. The first search was on "europe" and the second was on "north america." For each query, the user grouped the results.

Groups may be created using the Query Tool, or the items may be manually selected and then grouped, using the Group Tool. Groups are named sets of documents, which are assigned identifying colors for their Galaxies dots. Groups may be manipulated with standard set of operations, including Union, Intersection, and Exclusive-OR. In Figure 5, looking at the Group Tool panel, we can see the "europe" grouped results color-coded in pink and the "north america" results in red. An Intersection operation found five documents that occurred in both sets; these are marked in white and the dots have been replaced by 2's, indicating occurrence in two sets.

Together, the WebTheme visualizations and analysis tools enable users to quickly perceive the main themes within a collection of Web pages, locate pages relevant to the topic of interest, and determine where to spend analysis time.

V. CURRENT STATUS AND FUTURE WORK

WebTheme prototypes are now in use at PNNL and NASA. Certain users have specific interests in characterizing web sites, to get a quick-look at what information is contained in that web space, and then follow up with more in-depth analysis. These users begin the harvest from a list of URLs. In this case, the harvest is commonly restricted to the domains specified in the initial list of URLs, but the links are followed to greater depth. Users interested in analyzing results of a search typically begin by visiting the Alta Vista or Google search site and perform repeated searches to refine the query itself (as opposed to understanding the results). When that process is complete, the query string is provided to WebTheme, along with choices about how the harvest will be conducted. The depth of harvesting for these instances is typically shallow, as the linked information sources often quickly digress from the topic of interest.

All current users of WebTheme also have access to the full functionality of SPIRE, which offers more analytical tools for examining a collection. Thus, users who find a collection of special interest often move the set into SPIRE to complete the analysis.

WebTheme has already shown significant value for users interested in creating a visual representation of the conceptual content in a collection of Web documents, whether from a single site or the results of a search. However, because it is currently a beta prototype, several functions within the current version of WebTheme require additional research and development. These include additional functionality to the harvesting software, more intuitive user interface controls, and refined client visualization displays.

REFERENCES

1. Olsen, K. A., Korfhage, R. R., Sochats, K.M., Spring, M. B, & Williams, J. G. Visualization of a document collection: The VIBE system. Information Processing and Management, 29, 1 (1993) 69-81.

2. Hearst, M. A. TileBars: Visualization of Term Distribution Information in Full Text Information Access, in Proceedings of CHI '95 (Denver, Colorado, May 7-11, 1995) pp. 59-66.

3. IBM Alphaworks. 1999. "Mappucino". http://www.alphaworks.ibm.com/tech/mapuccino.

4. Battelle Memorial Institute. 2001. "SPIRE - Spatial Paradigm for Information Retrieval and Exploration" http://www.pnl.gov/infoviz/spire/spire.html .